InfiniBand™ Architecture Specification Volume 1

Release 1.0

October 24, 2000
Final

Certain IBA operations are valid only over certain classes of service. A QP rejects a WQE for an operation that is not valid for the configured class of service.

Connection oriented service requires that the consumer initiate a communication establishment procedure (connection setup) with the target node to associate the QPs and establish QP context prior to any QP operation. Actually, all service classes except for raw datagram need some form of communication setup to associate queue pairs. For reliable datagram service, the node performs a communication establishment process to associate an end-to end (EE) context (explained later) with each target node. All QPs configured for Reliable Datagram service use established EE contexts and the work request specifies which EE context to use for that operation.

Raw Datagrams are similar to unreliable datagrams, except that the source QP does not know the identity of the QP that will receive and process the message. Raw datagrams allow for routers that forward raw datagram packets to non IBA destinations on a disparate fabric (such as a LAN or WAN) that has no equivalent of a QP. There are two types of raw datagrams, IPv6 and Ethertype. IPv6 raw datagrams contain a global routing header and the packet payload contains a transport protocol service data unit as identified in the global routing header. An Ethertype raw datagram contains an Ethernet Type field and the packet payload contains a transport protocol service data unit as identified in the Ethernet Type field.

IBA defines both channel (send/receive) and memory (RDMA) semantics. Raw datagram and Unreliable Datagram services do not support memory semantics.

3.5.3 KEYS

IBA uses various keys to provide isolation and protection. Keys are values assigned by an administrative entity that are used in messages in various ways. The keys themselves do not provide security since the keys are available in messages that cross the fabric and thus any entity that can get to the interior of the fabric can ascertain key values. IBA does place restrictions on how applications can access certain keys.

The keys are:

- **Management Key** (M_Key): Enforces the control of a master subnet manager. Administered by the subnet manager and used in certain subnet management packets. Each channel adapter port has a M_Key that the SM sets and then enables. The SM may assign a different key to each port. Once enabled, the port rejects certain management packets that do not contain the programmed M_Key. Thus

only a SM with the programed M_Key can alter a node's fabric configuration. The SM can prevent the port's M_Key from being read as long as the SM is active. The port maintains a time-out such that the port reverts to an unmanaged state if the SM fails. There is one M_Key for a switch.

- **Baseboard Management Key** (B_Key): Enforces the control of a subnet baseboard manager. Administered by the subnet baseboard manager and used in certain MADs. Each channel adapter port has a B_Key that the baseboard manager sets. The baseboard manager may assign a different key to each port. Once enabled, the port rejects certain management packets that do not contain the programmed B_Key. Thus only a baseboard manager with the programed B_Key can alter a node's baseboard configuration. The baseboard manager can prevent the port's B_Key from being read as long as the baseboard manager is active. The port maintains a time-out such that the port reverts to an unmanaged state if the baseboard manager fails. There is one B_Key for a switch.
- **Partition Key** (P_Key): Enforces membership. Administered through the subnet manager by the partition manager (PM). Each channel adapter port contains a table of partition keys which is setup by the PM. QPs are required to be configured for the same partition to communicate (except QP0, QP1, and ports configured for raw datagrams) and thus the P_Key is carried in every IB transport packet. Part of the communication establishment process determines which P_Key that a particular QP or EEC uses. An EEC contains the P_Key for Reliable Datagram service and a QP context contains the P_Key for the other IBA transport types. The P_Key in the QP or EEC is placed in each packet sent, and compared with the P_Key in each packet received. Received packets whose P_Key comparison fails are rejected. Each switch has one P_Key table for management messages and may optionally support partition enforcement tables that filter packets based on their P_Key.
- **Queue Key** (Q_Key): Enforces access rights for reliable and unreliable datagram service (RAW datagram service type not included). Administered by the channel adapter. During communication establishment for datagram service, nodes exchange Q_Keys for particular queue pairs and a node uses the value it was passed for a remote QP in all packets it sends to that remote QP. Likewise, the remote node uses the Q_Key it was provided. Receipt of a packet with a different Q_Key than the one the node provided to the remote queue pair means that packet is not valid and thus rejected.

  Q_Keys with the most significant bit set are considered controlled Q_Keys (such as the GSI Q_Key) and a HCA does not allow a consumer to arbitrarily specify a controlled Q_Key. An attempt to send a controlled Q_Key results in using the Q_Key in the QP context. Thus

the OS maintains control since it can configure the QP context for the controlled Q_Key for privileged consumers.

- **Memory Keys** (L_Key and R_Key): Enables the use of virtual addresses and provides the consumer with a mechanism to control access to its memory. These keys are administered by the channel adapter through a registration process. The consumer registers a region of memory with the channel adapter and receives an L_Key and R_Key. The consumer uses the L_Key in work requests to describe local memory to the QP and passes the R_Key to a remote consumer for use in RDMA operations. When a consumer queues up a RDMA operation it specifies the R_Key passed to it from the remote consumer and the R_Key is included in the RDMA request packet to the original channel adapter. The R_Key validates the sender's right to access the destination's memory and provides the destination channel adapter with the means to translate the virtual address to a physical address.

3.5.4 VIRTUAL MEMORY ADDRESSES

IBA is optimized for virtual addressing. That is, an IBA consumer uses virtual addresses in work requests and the channel adapter is able to convert the virtual address to physical address as necessary. For this to happen, each consumer registers regions of virtual memory with the channel adapter and the channel adapter returns 2 memory handles called L_Key and R_Key to the consumer. The consumer then uses the L_key in each work request that requires a memory access to that region. See 3.5.3 for description of L_Key usage.

Memory Registration provides mechanisms that allow IBA consumers to de-scribe a set of virtually contiguous memory locations or a set of physically contiguous memory locations to allow the HCA to access the memory as a virtually contiguous buffer using virtual addresses.

IBA also supports remote memory access (RDMA) that permits a remote consumer to access that registered memory. For RDMA, the consumer passes the R_KEY and a virtual address of a buffer in that memory region to another consumer. That remote consumer supplies that R_Key in its RDMA WQEs that will access memory in the original node. See 3.5.3 for detailed description of R_Key usage.

3.5.5 PROTECTION DOMAINS

Not only does memory registration allow the use of virtual memory addressing, but it also provides an increased level of protection against inadvertent and unauthorized access.

Since a consumer might communicate with many different destinations but not wish to let all those destinations have the same access to its registered memory, IBA provides protection domains. Protection domains

node. After the operation is completed, the I/O unit then uses channel semantics to push an I/O completion message back to the processor node.

3.6.1 COMMUNICATION INTERFACE

"*Channel adapter*" is the term that identifies the hardware that connects a node to the IBA fabric (and includes any supporting software). The channel adapter for a processor node is called a "*host channel adapter*" (HCA) and a channel adapter in an I/O node is a "*target channel adapter*" (TCA). A consumer communicates through one or more "queue pairs" (QP). An HCA typically supports hundreds or thousands of QPs while a TCA might support less than ten QPs.

It is the QP that is the communication interface. The user initiates work requests (WR) that causes work items, called WQEs, to be placed onto the queues and the channel adapter executes the work item.

Specifically, the operations supported for Send Queues are:

- **Send Buffer** -- a channel semantic operation to push a local buffer to a remote QP's receive buffer. The Send WR includes a gather list to combine data from several virtually contiguous local buffer segments into a single message that is pushed to a remote QP's Receive Buffer. The local buffer's virtual addresses must be in the address space of the consumer that created the local QP.
- **RDMA Read** -- a memory semantic operation to read a virtually contiguous buffer on a remote node. The RDMA Read operation reads a virtually contiguous buffer on a remote endnode and writes the data to a local memory buffer.

  Like the Send operation, the local buffer must be in the address space of the consumer that created the local QP.

  The remote buffer must be in the address space of the remote consumer owning the remote QP targeted by the RDMA Read.
- **RDMA Write** -- a memory semantic operation to write a virtually contiguous buffer on a remote node. The WR contains a gather list of local buffer segments and the virtual address of the remote buffer into which the data from the local buffer segments are written.

  Like the Send WR, the local buffer must be in the address space of the consumer that created the local QP.

  The remote buffer must be in the address space of the remote consumer owning the remote QP targeted by the RDMA Write.
- **Atomic** -- a memory semantic operation to do an atomic operation on a remote 64 bit word. The Atomic operation is a combined Read, Modify, and Write operation.